DOCUMENT RESUME

ED 352 390 TM 019 294

AUTHOR Krippendorff, Klaus
TITLE Recent Developments in Reliability Analysis.
PUB DATE May 92
NOTE 22p.; Paper presented at the Annual Meeting of the International Communication Association (Miami, FL, May 21-25, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Definitions; Equations (Mathematics); *Mathematical Models; *Reliability; *Research Methodology; *Sample Size; Standards
IDENTIFIERS *Standard Setting

ABSTRACT

When one wants to set data reliability standards for a class of scientific inquiries or when one needs to compare and select among many different kinds of data with reliabilities that are crucial to a particular research undertaking, then one needs a single reliability coefficient that is adaptable to all or most situations. Work toward this goal from 1967 through the present is discussed. The agreement coefficient alpha has been defined as zero when the observed disagreement equals the disagreement that would be expected under conditions of chance, one when observed disagreement is absent, and negative when the observed exceeds the expected disagreement. A more recent definition corrects for small sample sizes, as an equation demonstrates. The route by which the equation was derived is traced to show generalization to: (1) any metric; (2) "m" observations or instruments; (3) missing data; (4) multiple values; and (5) many variables. Computer implementation and testing are planned to refine this versatile analytical device for reliability analysis in content analysis, survey research, and other data-generating procedures. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED352390

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

KLAUS KRIPPENDORFF

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Recent Developments in Reliability Analysis

by

Klaus Krippendorff
The Annenberg School for Communication
University of Pennsylvania

Paper prepared for presentation at the meeting of the
International Communication Association
in Miami FL, May 21-25, 1992.
Not for reproduction or quotation without author's permission.
Comments and criticism are appreciated.

TM019294

BEST COPY AVAILABLE

Recent Developments in Reliability Analysis

For many researchers, the literature of reliability coefficients seems bewildering although the methodological problem in which they are embedded is reasonably clear:

Since we can never know what it is that we claim to see independent of our seeing it, or, translated into the language of science, since we can not test hypotheses about reality without first generating the observations or data to talk about, the accuracy by which primary data "represent" an unobserved nature remains unascertainable in principle (Krippendorff, 1991). Yet, to assure that the data that go into a scientific inquiry are not accidental, it is important to demonstrate that the data-generating procedure is reproducible under varying circumstances and by several observers. All reliability measures are intended to express the degree to which several observers, several measuring instruments or several interrogations of the same units of analysis yield the same descriptive accounts, category assignments, quantitative measures or data for short.

But,

* reliability coefficients are often specialized to different metrics (levels of measurement). There are nominal scale coefficients, Scott's (1955) pi, for example, and interval scale coefficients, Kuder and Richardson's (1937) Formula #20, for example, that differ in the metric

to which they claim applicability but moreover stem from incompatible analytical traditions.

* Reliability coefficients have built-in assumptions that do not easily reveal themselves to users. For nominal scales alone, there is the measure of % (percent) agreement, there is Bennett, Alpert and Goldstein's (1954) S, there is Goodman and Kruskal's (1954) family of lambda coefficients, there is Scott's (1955) pi, there is Cohen's (1960) kappa, there is Perreault and Leigh's (1989) I and many more, not to forget my own alpha (Krippendorff 1967, 1970, 1980). The researcher will have difficulties to choose among them without detailed examination of their assumptions. Often this is not obvious. For example, Cohen's (1960) frequently used kappa turns out to be a hybrid that behaves like an agreement coefficient near its value of plus one and like an association (or correlation) coefficient near its value of zero and is hence not consistently interpretable between zero and one (Krippendorff, 1973) which are precisely the values important for reliability assessments.

* Reliability coefficients often vary in their ranges of values. Some range from zero to one, the notorious % agreement, for example, but also the proportion of systematic to total variance (Kuder and Richardson, 1937). Some range from minus to plus one (Scott, 1955; Cohen, 1960). Variations in their ranges make it virtually impossible to assign uniform meanings to the numbers they produce.

While one can always find reasons for preferring one coefficient over another, when one wants to set data reliability standards for a class of scientific inquiries, or when one needs to compare and select among many different kinds of data whose reliabilities are crucial to a particular research undertaking, then one needs a single coefficient adaptable to all or most situations.

To accomplish this, I have pursued a single simple form and defined the agreement coefficient alpha (Krippendorff, 1967, 1970, 1978, 1980) as

$$\alpha = 1 - \frac{D_{observed}}{D_{expected}} \qquad (1)$$

which is zero when the observed disagreement equals the disagreement that would be expected under conditions of chance, one when observed disagreement is absent and negative when the observed exceeds the expected disagreement--which can arise only under conditions of a consensus to disagree. While the plus one and zero values of this coefficient make alpha easily interpretable, the general form of (1) is common to several other coefficients as well and not yet specific about the assumptions that go into the definitions of the observed and the expected disagreements.

I said above that one can not state anything about reality until after data are made available and that reliability is or should be a measure of reproducibility,

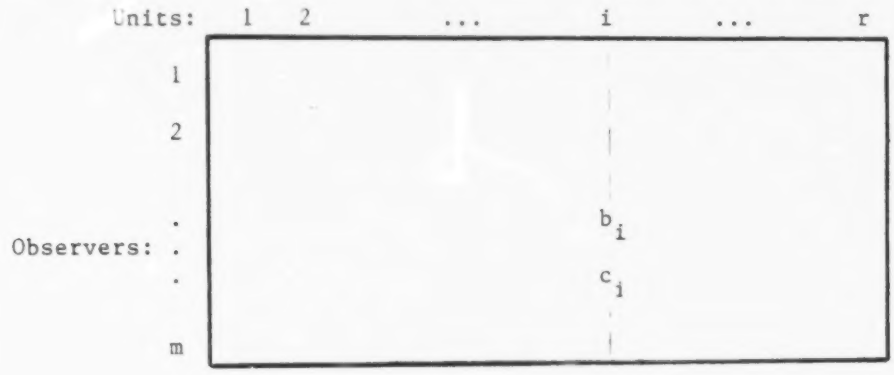
i.e., assess the result of applying a battery of claimed to be the same observational, accounting or measuring procedures to the same set of units of analysis. Under these conditions and without privileging one procedure over another, the only defendable assumption one can make about the "true nature" of the data relies on what all observers say they see or on what all measuring devices do indeed indicate. This assumption applies to both, the comparison of observational accounts or measures of any one unit as well as the characterization of the statistical distribution of the data to which all observers or measuring instruments jointly contribute. The latter is important for the definition of expected disagreement because it distinguishes measures of correlation or association (where correlates are known) from measures of agreement (Krippendorff, 1978) which must leave the nature of what is being observed uncharted.

There was another decision for alpha to correct for small sample sizes (small numbers of either units of analysis or observers/instruments or both). Many of the coefficients, particularly in content analysis, Scott's (1955) pi, for example, did not do this and systematically underestimated reliability when samples were small.

Let me start here with the result of my recent exploration and then show how I got there. I now define the general form of the agreement coefficient alpha by

$$\alpha = 1 - \frac{\sum_i \frac{n_i}{n_*} \bar{D}_i}{\bar{D}} \qquad (2)$$

where, in terms of the following canonical form of reliability data, an r-by-m table of n possibly quite unequally distributed observations, each generically denoted by b or c



the expected disagreement is

$$\bar{D} = \frac{1}{n(n-1)} \sum_b \sum_c d^2_{bc} \tag{3}$$

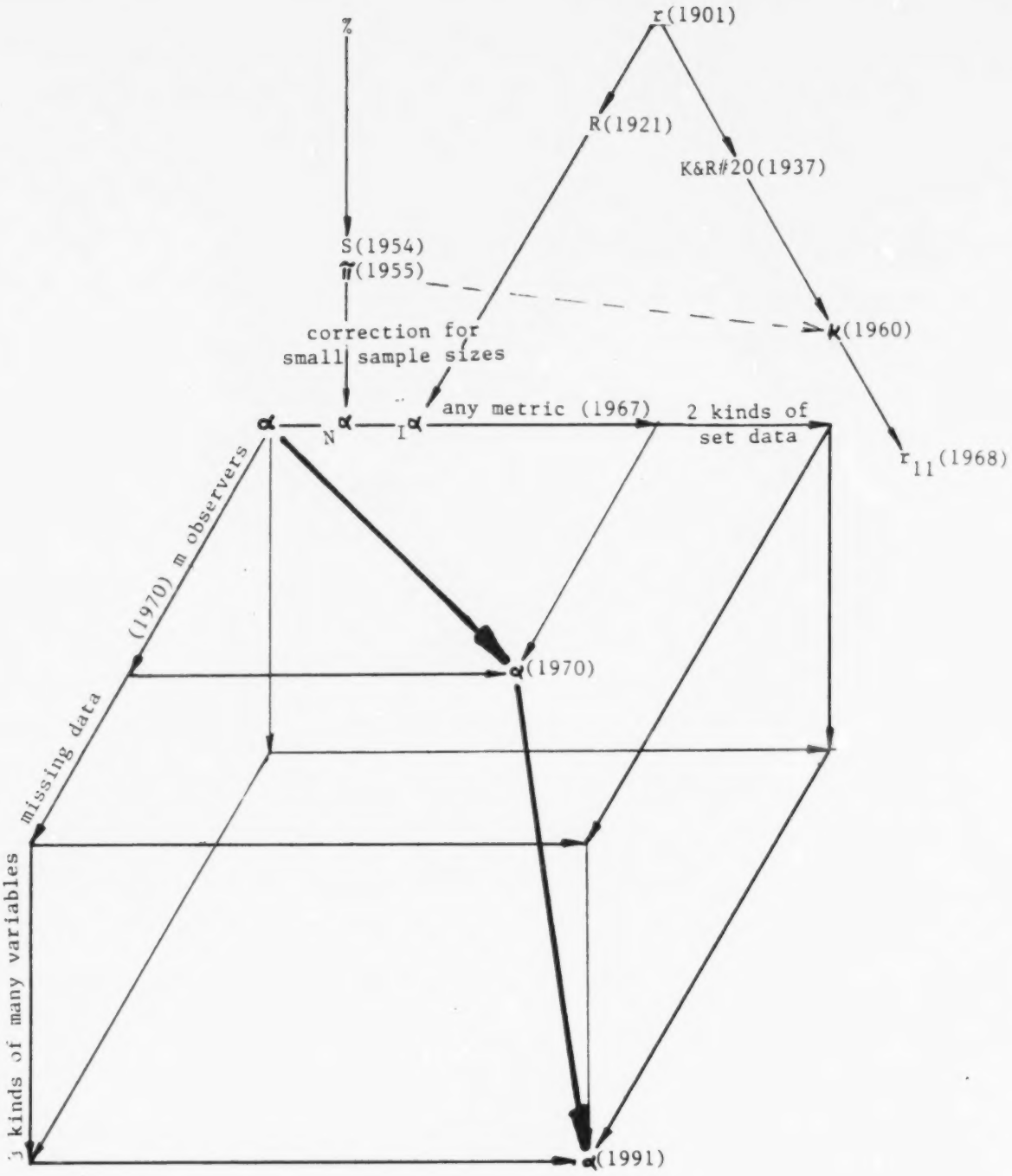
the disagreement within one unit i is

$$\bar{D}_i = \frac{1}{n_i(n_i-1)} \sum_{b_i} \sum_{c_i} d^2_{b_i c_i} \tag{4}$$

the number of observations that actually do contribute to any comparison within units is

$$n_\star = \sum_i n_i \,\Big|\, n_i > 1 \tag{5}$$

$$n_\star \leq n \leq rm \tag{6}$$



Creating a Larger Space for the Agreement Coefficient Alpha

Figure 1

and d_{bc} is a difference between any two observations, values or data points.

Before going into various manifestations of alpha, let me introduce with Figure 1 a kind of travel plan that shows how my thinking developed and charts the space within which alpha is applicable with comparable results.

Figure 1

Generalization to any metric was accomplished well before 1967 when I wrote a computer program using the following difference functions (the first four are also found in Table 1) each associated with one scale of measurement.

$$\text{nominal} \quad {}_N d^2_{bc} = \begin{cases} 1 \text{ iff } b \neq c \\ 0 \text{ iff } b = c \end{cases} \tag{7}$$

$$\text{ordinal} \quad {}_O d^2_{bc} = \left(\sum_{k<b} n_k + \frac{n_b - n_c}{2} - \sum_{k<c} n_k \right)^2 \tag{8}$$

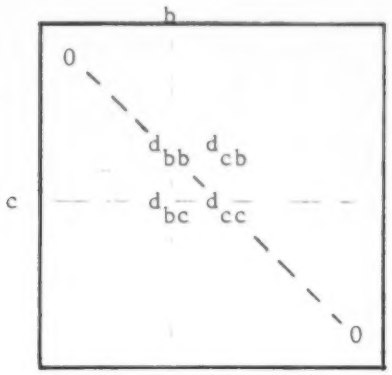
$$\text{interval} \quad {}_I d^2_{bc} = (b-c)^2 \tag{9}$$

$$\text{ratio} \quad {}_R d^2_{bc} = \left(\frac{b-c}{b+c} \right)^2 \tag{10}$$

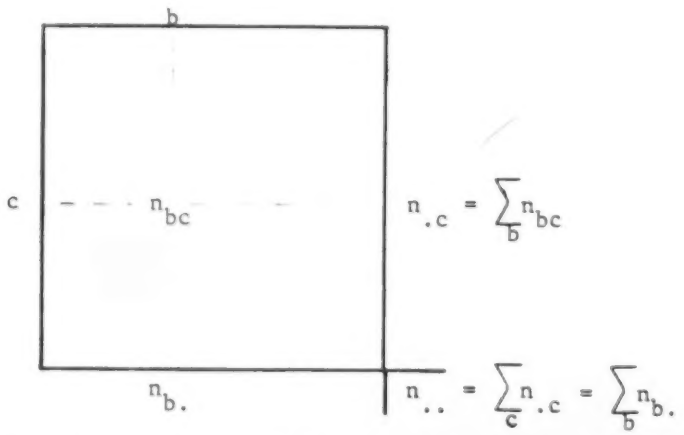
$$\text{polar} \quad {}_P d^2_{bc} = \frac{(b-c)^2}{(b+c-2k_{min})(2k_{max}-b-c)} \tag{11}$$

In (3) and (4) one can see that measures of disagreement enumerate all possible pairs characterized by these differences and in (12) one can see that these

difference functions serve as weights of the frequency of pairs of observations. One can also enter them in a metric-specific matrix of all possible pairs of values whose diagonal, containing all perfect matches, is always zero.



Originally, I was thinking only of two observers, coders or measuring instruments and as is customary in the literature in terms of <u>contingency</u> matrices that contain one pair of values for each unit of analysis, $r=n_{..}$ in number.



For two observers and in terms of the above contingency matrix notations:

$$\alpha = 1 - \frac{\sum_b \sum_c n_{bc} d_{bc}^2}{\sum_b \sum_c e_{bc} d_{bc}^2} \qquad (12)$$

Wherein the expected frequencies are

$$e_{bc} = \frac{(n_{b.} + n_{.b})(n_{c.} + n_{.c} - \Delta_{bc})}{4n_{..} - 2} \qquad (13)$$

and the Kronecker delta is defined by

$$\Delta_{bc} = \begin{cases} 1 \text{ iff } b=c \\ 0 \text{ iff } b\neq c \end{cases} \qquad (14)$$

The two observer nominal scale version of alpha in (12) became Scott's (1955) pi, corrected for small sample sizes, the relation between the two being

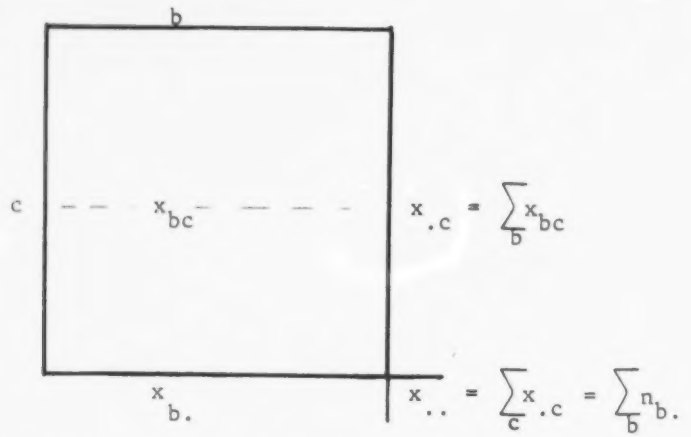
$$\alpha = 1 - \frac{2n_{..} - 1}{2n_{..}} \tilde{\pi} \qquad (15)$$

and the interval scale version of (12) became Pearson's (1901; Tildesley, 1921) intra-class correlation coefficient R. All five versions of alpha are obtained by inserting the difference functions (7) through (11) into (12). They assume the same value for bivariate decisions, i.e., for 2-by-2 contingency tables, and for coincidence matrices whose off-diagonal entries are in exact proportion to the expected frequencies. Thus, a comparison of the coefficients computed with different difference functions could also

reveal the nature of the metric most likely underlying the data at hand.

Generalization to m observers/instruments required a measure of disagreement applicable to more complex patterns of differences among observers. I adopted a measure that aggregates all pairwise differences obtainable within a set of observations (see (3) and (4)). This is by no means the only function possible. Entropy measures would do much the same at least for nominal data but my choice was to conform to the conventions of most other available statistical techniques, particularly in the correlational or analysis of variance tradition, to which data are likely subjected once reliability was established. Indeed, one can argue that reliability should ideally reflect the disagreements that matter in subsequent analyses and these can often be reduced to aggregates of pairwise differences.

In terms of m-dimensional contingency matrices the conception of aggregate differences proved difficult to state. I therefore moved from contingency to coincidence matrix representations of reliability data (Krippendorff, 1980) whose notations do not separate the individual observer's contributions to these data. A coincidence matrix of all possible pairs observed within units



contains these entries (n_{b_i} being the number of b's in unit i)

$$x_{bc} = \sum_{i=1}^{r} n_{b_i}(n_{c_i} - \Delta_{bc}) \tag{16}$$

and enables one to state <u>alpha</u> as

$$\alpha = 1 - (x_{..} - m + 1)\frac{\sum_b \sum_c x_{bc} d_{bc}^2}{\sum_b \sum_c x_{b.} x_{.c} d_{bc}^2} \tag{17}$$

<u>Generalization to missing data</u> turned out to be a natural extension of the generalization to m observers but could no longer rely on coincidence matrix notions. When some of the rm possible entries in the canonical form of reliability data are blank, there are therefore no longer rm observations but a number n smaller than rm. Here, special attention had to be paid to the fact that under these conditions only units with two or more observations can contribute to the observed disagreement which defined a number n_* (5) that excludes observations in units with no or only one entry. By inserting (3) and (4) into (2) which are stated here in terms of the canonical form of

reliability data and without computational shortcuts, the agreement coefficient alpha then takes its most general form

$$\alpha = 1 - \frac{n(n-1)}{n_*} \frac{\sum_i \frac{1}{n_i - 1} \sum_{b_i} \sum_{c_i} d^2_{b_i c_i}}{\sum_b \sum_c d^2_{bc}} \qquad (18)$$

Generalization to multiple values. When each unit of analysis is assigned exactly one value by each observer, then any observed difference d_{bc} between the two values b and c two observers respectively assign to one unit must contribute to the unreliability of data. However, it may happen that observers are asked to represent each unit by an appropriate set of descriptors (keywords of articles, lists of attributes, multiple category assignments, etc.). Here, differences within any such set must not contribute to unreliability. The problem therefore is to define a difference function between two sets B and C where pairwise differences within B and within C are ignored while pairwise differences across B and C are aggregated appropriately.

Differences between two sets B and C can be conceptualized in two ways. The core difference, i.e. the difference between the most representative element of each set (modes, medians, means, mean products, etc., depending on the metric involved) makes no reference to the size or spread of values in the set. The set difference seeks to match the values in one set with those in the other, ideally on a one-to-one basis, and thus responds to differences in

the sets' magnitudes and distribution. Since core differences are relatively transparent, they are simply listed in Table 1. The set-difference function proved problematic for a long time. The solution here proposed came about in conjunction with work on a general calculus for disagreement analysis (in progress). Without going into detailed justification, given a set W

$$W = \{w_1, w_2, \ldots, w_{n_W}\}$$

disagreement within W is

$$D_W = \frac{1}{n_W} \sum_{i=1}^{n_W} \sum_{j=1}^{n_W} d^2_{w_i w_j} \qquad (19)$$

and disagreement within the union of two sets B and C (with duplications included) is

$$D_{B \cup C} = \frac{1}{n_B + n_C} \sum_b \sum_c d^2_{bc} \qquad (20)$$

This distance function between sets B and C now becomes

$$d^2_{BC} = D_{B \cup C} - \max_{W=B,C} \left(\frac{n_B + n_C}{n_W} D_W \right) \qquad (21)$$

This difference function is zero when the two sets have the same values and in the same proportion and increases with any deviation from the ideal of a one-to-one correspondence. Such a difference of differences is simply

Metric	Two Values (Same Metric) b c Simple Differences	Two Sets of Values (Same Metric) $B = \{b_1, b_2, \ldots, b_j, \ldots, b_r\}$; $r = n_B$ $C = \{c_1, c_2, \ldots, c_j, \ldots, c_s\}$; $s = n_C$ Core Differences	Set Differences	Multivariate Differences	Maximum Differences
Nominal	(.1) $N^{d}_{bc} = (1 - \Delta_{bc}) = \begin{cases} 0 \text{ iff } b=c \\ 1 \text{ iff } b \neq c \end{cases}$	(1.1) $\tilde{b}$ = mode of B $\tilde{c}$ = mode of C $N^{d}_{\tilde{b}\tilde{c}} = 1 - \frac{\#(\tilde{b} \cap \tilde{c})}{\#(\tilde{b} \cup \tilde{c})}$	(2.) $D_B = \frac{1}{r} \sum_{i=1}^{r} \sum_{j=1}^{r} d^2_{b_i b_j}$ $D_C = \frac{1}{s} \sum_{i=1}^{s} \sum_{j=1}^{s} d^2_{c_i c_j}$ let $b, c \in B \cup C$ $D_{B \cup C} = \frac{1}{r+s} \sum_{b} \sum_{c} d^2_{bc}$ $d_{BC} = \sqrt{D_{B \cup C} - \max_{w \in B,C} \left(\frac{n_B + n_C}{n_w} D_w \right)}$	(3.) $d_{\langle b \rangle \langle c \rangle} = \sqrt{\sum_{j=1}^{m} v_j \frac{d^2_{b_j c_j}}{D_j}}$	(4.) $d_{\max b_j c_j} = \sqrt{\max_{j=1}^{m} \left(v_j \frac{d^2_{b_j c_j}}{D_j} \right)}$
Ordinal	(.2) $O^{d}_{bc} = \frac{n_b}{2} + \sum_{k<b} n_k - \frac{n_c}{2} - \sum_{k<c} n_k$	(1.2) $\tilde{b}$ = median of B $\tilde{c}$ = median of C $O^{d}_{\tilde{b}\tilde{c}} = \frac{n_{\tilde{b}}}{2} + \sum_{k<\tilde{b}} n_k - \frac{n_{\tilde{c}}}{2} - \sum_{k<\tilde{c}} n_k$			
Interval	(.3) $I^{d}_{bc} = b - c$	(1.3) $\bar{b} = \frac{1}{r} \sum_{j=1}^{r} b_j$ = mean b $\bar{c} = \frac{1}{s} \sum_{j=1}^{s} c_j$ = mean c $I^{d}_{\bar{b}\bar{c}} = \bar{b} - \bar{c}$			
Ratio	(.4) $R^{d}_{bc} = \frac{b - c}{b + c}$	(1.4) $\hat{b} = \sqrt{\frac{1}{r} \sum_{j=1}^{r} b_j^2}$ $\hat{c} = \sqrt{\frac{1}{s} \sum_{j=1}^{s} c_j^2}$ $R^{d}_{\hat{b}\hat{c}} = \frac{\hat{b} - \hat{c}}{\hat{b} + \hat{c}}$			

BEST COPY AVAILABLE

A Comparison of Difference Functions

Figure 1

inserted in the expressions for the observed and expected disagreements (3), (4), and alpha (12), (17), (18) and in forms generated by the following generalizations to many variables.

Generalizations to many variables can be treated as extensions of the idea of finding differences of differences, not merely between two sets of values (which one can also interpret as the values of two unordered sets of variables), but also within multivariate spaces or between profiles.

Aggregating the reliabilities of several variables into a one quantity is important whenever one wants to compute a single indicator from a set of possibly unreliable measures or one wants to test a hypothesis on several such variables and the reliabilities of these variables are known to stand in a certain relation to each other. We distinguish four kinds whose computational consequences are listed in Table 1

Table 1

(1) When variables are averaged across variables in a subsequent analytical step and the disagreements in one variable could compensate for disagreements in another, then the above mentioned core differences are suggested as an appropriate form of aggregation.

(2) When variables are taken as unordered sets whose values must in some form be unscrambled or matched in subsequent analyses, then the above mentioned set

differences are suggested as an appropriate form of aggregation.

(3) When the values of variables are fixed in number and ordered, i.e. designate a point in a multivariate space, then a form of the hyper geometric difference between any two such points is required. I have opted to take the Mahalanobis (1936) distance as a starting point for this multivariate distance function because it corrects for the magnitude of variation in each variable, giving each variable the same weight (this can be altered, of course, by assigning a weight v depending on how the variables in question contribute to subsequent analyses). This form of aggregation is particularly appropriate when correlation or variance analyses are intended to be performed on these data.

A side advantage of this multivariate distance function is that the variables in question may be of entirely different metric (nominal, ordinal, interval, ratio-variables) or consist of unordered sets of values of any one kind.

Although it is tempting to use this form of aggregation to evaluate the reliability of a whole multi-variate measuring instrument, the results may be deceptive because many highly reliable variables can easily blank out a few very unreliable variables and lead to unwarranted acceptance of data.

(4) When variables cannot be considered orthogonal and geometrically compensatory of each other's reliabilities, maximum differences provide an appropriate form of aggregation. Here the maximum of any simple difference between the values of two arrays of ordered values that are being compared is taken to be the aggregate difference. Maximum differences provide the most severe criteria for the reliability in multivariate spaces, core differences the least.

The work reported here is still in progress. Computer implementation and testing is planned. The hope is to create an extremely versatile analytical device for the analysis of the reliabilities in content analysis, survey research and a variety of other data generating procedures.

References

Bennett, E.M., Alpert, R. and Goldstein, A.C. (1954). Communication Through Limited Response Questioning. *Public Opinion Quarterly*, 18, 303-308.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Bulletin*, 76, 378-382.

Goodman, Leo A., and Kruskal, William H. (1954). Measures of Association for Cross Classification. *Journal of the American Statistical Association*, 49, 733-764.

Krippendorff, Klaus (1967). A Computer Program for Analyzing Multivariate Agreements, User's Manual. Philadelphia: The Annenberg School of Communications, University of Pennsylvania, Mimeo [Version 2, 1970; Version 3, 1972].

Krippendorff, Klaus (1970). Bivariate Agreement Coefficients for Reliability of Data. In Borgotta, E.R. and Bohrnstedt, G.W., (Eds.) *Sociological Methodology 1970* Vol. 2. San Francisco, CA: Jossey-Bass, Inc.

Krippendorff, Klaus (1978). Reliability of Binary Attribute Data. *Biometrica*, 34, 142-144.

Krippendorff, Klaus (1980). *Content Analysis; An Introduction to its Methodology*. Beverly Hills, CA: Sage Publications.

Krippendorff, Klaus (1987). Association, Agreement and Equity. *Quality and Quantity*, 21, 109-123.

Krippendorff, Klaus (1991). Reconstructing (some) Communication Research Methods. In Steier, F. (Ed.) *Research and Reflexivity*. London: Sage Publications.

Kuder, G.F., and Richardson, M.W. (1937). The Theory and Estimation of Test Reliability. *Psychometrica*, 2, 151-160.

Mahalanobis, P.C. (1936). On the Generalized Distance in Statistics. *Proceedings of the National Institute of Science* (India), 12, 49-55.

Maxwell, A.E. and Pilliner, A.E.G. (1968). Deriving Coefficients of Reliability and Agreement for Ratings. *British Journal of Mathematical and Statistical Psychology*, 21, 105-116.

Pearson, K. (1901). Mathematical Contributions to the Theory of Evolution. IX: On the Principle of Homotyposis and its

Relation to Heredity, to Variability of the Individual, and to that of Race. Part I: Homotyposis in the Vegetable Kingdom. *Philosophical Transactions of the Royal Society* (London). Series A, 193, 358-479.

Perreault, W.D., and Leigh, L.E. (1989). Reliability of Nominal Data based on Qualitative Judgements. *Journal of Marketing Research*, 26, 135-148.

Scott, W. A. (1955). Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly* 19, 321-325.

Tildesley, M.L. (1921). A First Study of The Burmese Skull. *Biometrica*, 13, 176-267.